2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)
April 3-7, 2020, Iowa City, Iowa, USA

OCT IMAGE QUALITY EVALUATION BASED ON DEEP AND SHALLOW FEATURES FUSION NETWORK

Rui Wang[1], Dongyi Fan[1], Bin Lv[1], Min Wang[2], Qienyuan Zhou[3], Chuanfeng Lv[1], Guotong Xie[1], Lilong Wang[1]
[1]PingAn Technology (Shenzhen) Co., Ltd., Shenzhen, China
[2]Department of Ophthalmology, Eye and ENT Hospital of Fudan University, Shanghai, China
[3]Optovue Inc., Fremont, California, USA

ABSTRACT

Optical coherence tomography (OCT) has become an important tool for the diagnosis of retinal diseases, and image quality assessment on OCT images has considerable clinical significance for guaranteeing the accuracy of diagnosis by ophthalmologists. Traditional OCT image quality assessment is usually based on hand-crafted features including signal strength index and signal to noise ratio. These features only reflect a part of image quality, but cannot be seen as a full representation on image quality. Especially, there is no detailed description of OCT image quality so far. In this paper, we firstly define OCT image quality as three grades ('Good', 'Usable' and 'Poor'). Considering the diversity of image quality, we then propose a deep and shallow features fusion network (DSFF-Net) to conduct multiple label classification. The DSFF-Net combines deep and enhanced shallow features of OCT images to predict the image quality grade. The experimental results on a large OCT dataset show that our network obtains state-of-the-art performance, outperforming the other classical CNN networks.

Index Terms— Optical coherence tomography, Image quality assessment, Deep learning

1. INTRODUCTION

Optical coherence tomography (OCT) is a powerful imaging technique in the field of ophthalmology, which is similar to retinal histological image showing detailed retinal structures. It can detect the cross-sectional retinal structures and their pathological changes to help clinical diagnosis [1]. With the development of deep learning technology, more and more studies achieve retinal diseases screening and computer aided diagnosis by combining OCT imaging and deep learning technology. [2, 3]. OCT images with poor quality could lead to inaccurate results of these automatic analysis. While image quality is usually affected by ocular media opacity, eye diseases, as well as operator experience. Although modern imaging systems are less operator dependence than previous generation systems, operators still play a key role [4]. Therefore, reliable image quality assessment is important in OCT clinical use.

Rao et. al confirmed that the quality of OCT scans adversely affected the accuracy of glaucoma detection with retinal nerve fiber layer and ganglion cell complex parameters [5]. Tanga et. al studied the effect of pupil dilation affecting the quality of OCT imaging as well [6]. Folio et. al verified that the OCT scan quality could reduce the local retinal nerve fiber layer segmentation error [7]. These OCT image quality evaluation systems used in the above studies were based on hand-crafted features including the signal strength index (SSI). SSI measures the quality of the image signal in OCT imaging stage and reflects the quality of a volume image, rather than the quality of each image. In order to solve this problem, Stein et. al proposed a new OCT image quality index (QI) [8]. The SSI and SNR of the OCT image are taken into consideration in QI system and finally it is obtained by image histogram distribution [8]. The QI is compared with the scores of three OCT experts to prove the reliability of the results. However, this method only used 63 OCT images and lacked large-scale data validation.

To solve these problems, we firstly propose an image quality grading system (IQGS) for OCT images, which includes 'Good', 'Usable' and 'Poor'. Secondly, we design a new convolutional network called deep and shallow features fusion network (DSFF-Net) to grade OCT images. Finally, we evaluate our image quality assessment method on a large sample dataset with 12538 OCT images and compare with other classification networks.

2. METHODS

2.1. Image Quality Grading System

We propose an image quality grading system (IQGS) for OCT image. Based on criteria set up by clinical experts, we divide OCT images into three levels, which are 'Good', 'Usable' and 'Poor' in our IQGS. 'Good' corresponds to 'normal' image quality. 'Usable' corresponds to an image quality which is 'blur' or 'slight deficiency'. Critical features and lesions are still recognizable on those images. And 'Poor' corresponds to an image quality which is 'noise' or

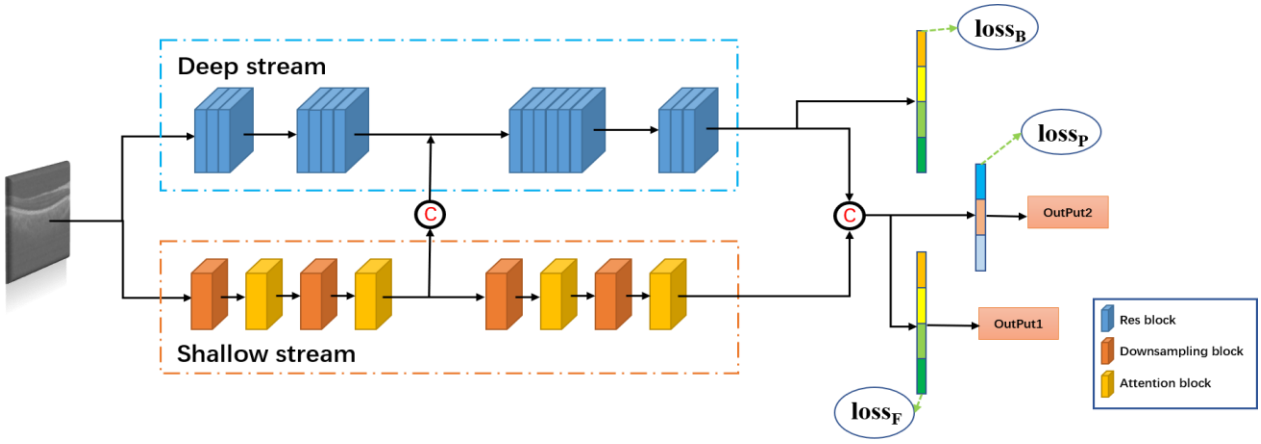

Fig. 1: The architecture of DSFF-Net. The deep stream backbone is ResNet50. And a down sampling module is used to re-size the input image and an attention module is employed to extract the low dimension feature within shallow stream. At last, a fusion block output the result of classification.

'deficiency' and the corresponding images cannot be used to provide a credible diagnosis.

2.2. Network Architecture

In traditional OCT image quality assessment system, SSI and SNR are usually used to evaluate OCT image quality. However, neither parameter is effective for detecting quality problems such as cropped OCT images, partial blink, blurriness, or regional weak signal. To solve this problem, we design a deep convolution neural network to extract image features for image quality assessment. At the prediction stage, the quality grade of each image is summarized through our network. The traditional convolution neural network can extract the deep abstract features of image by deepening network and synthesize these features to get the results, such as ResNet [9], DenseNet [10] and MobileNetV2 [11]. However, these networks often ignore the shallow features of the image that is the most critical information in the image.

DSFF-Net we propose integrates deep and shallow image features using ResNet50 as the backbone. The network structure is shown in Fig. 1. DSFF-Net has two streams, in which one is a deep feature extraction stream based on ResNet50 and the other is a shallow feature extraction stream. The shallow stream consists of a down-sampling block and an attention block. The down-sampling block of DSFF-Net contains a 3×3 convolutional layer with 2 strides, a batch-normalization layer and an activation layer. It extracts the shallow features when down-sampling the image. In order to focus on the feature area, we introduce the attention block which includes channel attention module and spatial attention module. For the channel attention, we use the max-pooled features along with the average-pooled features [12,13]. The results indicate that the combination enables the encoder to represent channel feature on global dimension, and enhances the effect of more salient channel. Thus, we exploit the mixed pooling feature by element-wise summation [13]. After the channel attention block, we implement the average pooling and max pooling likewise, and then compress the output which can be considered as the spatial feature attention map. Subsequently the attention map is element-wise multiplied to the original feature. At the output stage, 'Output1' is the prediction results of the four sub-labels and 'Output2' outputs final quality grading prediction results.

2.3. Loss Function

Unlike traditional classification networks, which only use the loss function of the last layer to train the whole model, DSFF-Net improves their accuracy of the classification results by retaining loss functions of the base networks, and combines all loss function of the last layer which can be written as:

$$Loss_{total} = w_B Loss_B + w_F Loss_F + w_P Loss_P \quad (1)$$

where $Loss_B$ represents the loss on backbone network, $Loss_F$ represents fusion loss and $Loss_P$ is the predictive loss. $Loss_B$ and $Loss_F$ are both binary-class cross entropy, because they are the loss function of multi-label classification branches. $Loss_P$, which is the objective function for the multi-class classification branch, is categorical cross entropy. w_B, w_F and w_P are weights for each part of aforementioned loss, and we set w_B=0.3, w_F=0.3, w_P=0.4 respectively.

3. EXPERIMENT

In this section, we conduct a validation experiment to investigate the effectiveness of our DSFF-Net on OCT image dataset.

3.1. Datasets

12538 OCT images are collected from Eye and ENT Hospital of Fudan University by using SD-OCT devices (iScan/iVue, Optovue), with the same intervals covering 7 × 7 mm range of macular region. The dataset is annotated by 2 ophthalmologists and the images are checked by a third clinical expert when differences occur. Among the labels, exclusive relation exists in the image quality which is good, usable (slight deficiency and blur) and poor (deficiency and

noise). A summary of this dataset is given in Table 1, and Fig. 2 shows some sample images of this dataset.

3.2. Training details

Here we consider the quality grading as a multi-label classification problem. Processes of resizing from 1024×640 to 512×320 and padding to 512×512 are applied to input images before training and inference period.

Table 1: Distribution of our OCT image dataset.

		Train	Test	All
Poor	Noise	1457	485	1942
	Deficiency	1832	610	2442
	Noise + Deficiency	686	228	914
Usable	Blur	1172	390	1562
	Slight deficiency	749	250	999
	Blur + Slight deficiency	483	162	645
Good	Normal	3026	1008	4034
	Total	9405	3133	**12538**

At the stage of image pre-processing, we create some blur images randomly by image fusion and Gaussian blur to relieve sample imbalance. For data augmentation, common practice is followed, including random horizontal flipping and random rotation. Since the SNR and the scanning quality vary on grayscale, color jittering is not considered so as to prevent unexpected noise. The model we propose is optimized using SGD optimizer with a learning rate of 0.001. The framework is implemented on Keras with Tensorflow backend.

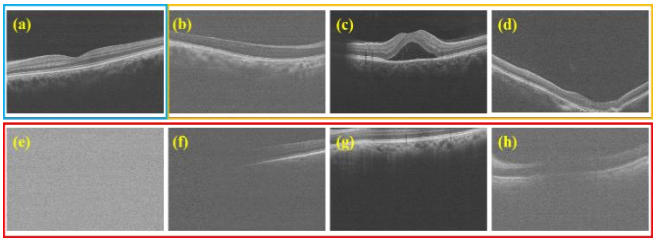

Fig.2: Examples of different OCT images quality grades. The images of 'good' quality (a) could provide effective diagnosis, the images between 'Poor' and 'Good' (b-d) have some poor-quality indicators, but it is not enough to affect diagnosis, and the 'poor' images (e-h) are incompetent for diagnosis.

3.3. Quantitative Evaluation

To evaluate our model's performance, we implement classification comparative experiments on various representative backbones including ResNet50 [9], DenseNet121 [10], and MobileNetV2 [11].

In order to analyze the performance of our method both on three grades and four fine-grained labels, we compute Area Under the ROC Curve (AUC) for each label on multi-label evaluation (4 labels), Accuracy, Precision and Recall on test subset on each model. Considered that models with different blocks affect performances in various streams, the evaluation is set by a variable-controlling approach. The factors contain streams, attention blocks and softmax-sigmoid combination. The AUCs of different labels are demonstrated in Table 2, and the performances of different models for image quality grading are reported in Table 3.

Table 2: The AUC of different labels in various network.

	Noise	Deficie-ncy	Blur	Slight Deficie-ncy	Overall
MobileNet-V2	0.9235	0.9153	0.9653	0.9100	0.9285
ResNet50	0.9862	0.9703	**0.9939**	0.8872	0.9594
DenseNet121	0.9765	0.9823	0.9472	0.9621	0.9670
DSFF-Net (No Attention)	**0.9963**	**0.9930**	0.9080	0.9800	0.9693
DSFF-Net	0.9950	0.9917	0.9925	**0.9884**	**0.9919**

As is shown in Table 2, the DSFF-Net performed relatively well on both 'deficiency' and 'slight deficiency' labels, because the two labels are more concerned with local features extracted from shallow networks. At the same time, the AUCs of other sub-labels also improve to some extent.

After the addition of fine-grained labels, the performance of multi-class classification improves obviously according to the Table 3. The fusion of multiple labels and classes which brings mutual exclusion provides significant gains. For ResNet50, we only use softmax-sigmoid combination activation structure achieves average accuracy at 0.9556, which is much higher than 0.9206 of the original ResNet50. Furthermore, our DSFF-Net with attention block obtains better performance, no matter on each class or in general, which obtains average accuracy at 0.9693 with DSFF-Net that does not contain attention block and 0.9746 with DSFF-Net containing attention block. Although the performance of DSFF-Net is not the best on all the evaluation indices of every grade such as those of 'Good', our model gains the highest score on the three assessment indices of overall results.

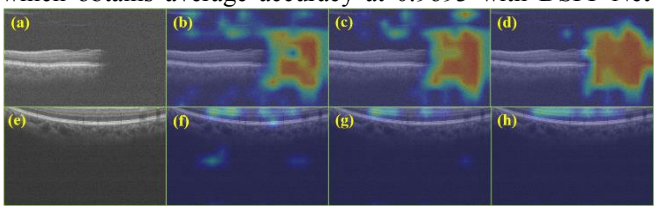

Fig.3: CAM heat map of various model on the sample labeled deficiency. Input images include (a) and (e), while (b)& (f), (c)&(g) and (d)&(h) are CAM output of deep feature stream, DSFF-Net (No attention) and DSFF-Net.

To verify the effectiveness and interpretability of DSFF-Net, we illustrate class activation mapping (CAM) [14] heat

Table 3: Performance comparison of quality classification methods test on different network. Mode-A adopts only multi-category classification method and Mode-B integrates both multi-class and multi-label classification method. P is precision, R is recall, and ACC means accuracy.

Model		Poor		Usable		Good		Overall		
		P	R	P	R	P	R	P	R	ACC
Mode-A	**MobilNetV2**	0.9148	0.9574	0.9371	0.8817	**0.9950**	0.9852	0.9490	0.9414	0.9460
	ResNet50	**0.9778**	0.8527	0.8441	0.9290	0.9269	**1.0000**	0.9163	0.9272	0.9206
	DenseNet121	0.9753	0.9186	0.9200	0.9527	0.9528	0.9951	0.9494	0.9554	0.9524
Mode-B	**MobilNetV2**	0.9298	0.8721	0.9474	0.8521	0.8559	0.9951	0.9110	0.9064	0.9063
	ResNet50	0.9502	0.9613	0.9286	0.9231	0.9851	0.9754	0.9546	0.9532	0.9556
	DenseNet121	0.9529	0.9419	0.9576	0.9349	0.9524	0.9852	0.9543	0.9540	0.9540
	DSFF-Net (No Attention)	0.9549	0.9845	0.9752	0.9290	0.9852	0.9852	0.9718	0.9662	0.9698
	DSFF-Net	0.9621	**0.9845**	**0.9758**	**0.9527**	0.9900	0.9803	**0.9760**	**0.9725**	**0.9746**

map for each model, as is shown in Fig. 3. It can be clearly seen that the masks of samples labeled deficiency focus on the discriminate region more, and the activation value of irrelevant region is suppressed remarkably. It proves that DSFF-Net performs better than other methods.

4. CONCLUSION

OCT image quality assessment plays a key role in area of medical imaging. There are two main contributions in this study. Firstly, we propose IQGS for OCT images quality assessment for the first time. It is a general grading system that can be applied on most OCT images. Secondly, we design a new convolutional network called deep and shallow features fusion network (DSFF-Net) to grade OCT images. We applied a series of experiments compared with various commonly used baseline models, including MobileNetV2, ResNet50 and DenseNet121. The experimental result reveals that our proposed DSFF-Net outperforms other existing methods in field of OCT image quality assessment.

5. REFERENCE

[1] Browning, David J., et al. "Comparison of the clinical diagnosis of diabetic macular edema with diagnosis by optical coherence tomography." Ophthalmology 111.4 (2004): 712-715.

[2] De Fauw, Jeffrey, et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease." Nature Medicine 24.9 (2018): 1342.

[3] Kermany, Daniel S., et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." Cell 172.5 (2018): 1122-1131.

[4] Hardin, Joshua S., et al. "Factors affecting Cirrus-HD OCT optic disc scan quality: a review with case examples." Journal of Ophthalmology 2015 (2015).

[5] Rao, Harsha L., et al. "Effect of scan quality on diagnostic accuracy of spectral-domain optical coherence tomography in glaucoma." American Journal of Ophthalmology 157.3 (2014): 719-727.

[6] Tanga, Lucia, et al. "Evaluating the effect of pupil dilation on spectral-domain optical coherence tomography measurements and their quality score." BMC Ophthalmology 15.1 (2015): 175.

[7] Folio, Lindsey S., et al. "Variation in optical coherence tomography signal quality as an indicator of retinal nerve fibre layer segmentation error." British Journal of Ophthalmology 96.4 (2012): 514-518.

[8] Stein, et al. "A new quality assessment parameter for optical coherence tomography." British Journal of Ophthalmology 90.2 (2006): 186-190.

[9] He, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[11] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[12] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[13] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision. 2018.

[14] Selvaraju, et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.